VEGETATION CLASSIFICATION AND SURVEY

International Association for Vegetation Science (IAVS)

Vegetation Classification and Survey 1: 77–86
doi: 10.3897/VCS/2020/48518

RESEARCH PAPER ECOINFORMATICS

Finite Mixture Model-based classification of a complex vegetation system

Fabio Attorre[1], Vito E. Cambria[2], Emiliano Agrillo[3], Nicola Alessi[4], Marco Alfò[5], Michele De Sanctis[1], Luca Malatesta[1], Tommaso Sitzia[2], Riccardo Guarino[6], Corrado Marcenò[7], Marco Massimi[1], Francesco Spada[1], Giuliano Fanelli[1]

1 *Department of Environmental Biology, Sapienza University of Rome, Italy*
2 *Department of Land, Environment, Agriculture and Forestry, University of Padova, Legnaro, Italy*
3 *Institute for Environmental Protection and Research (ISPRA), Rome, Italy*
4 *Faculty of Science and Technology, Free University of Bozen-Bolzano, Italy*
5 *Department of Statistical Sciences, Sapienza University of Rome, Italy*
6 *Department of Environmental Biology and Biodiversity, University of Palermo, Italy*
7 *Department of Plant Biology and Ecology, University of the Basque Country, Bilbao, Spain*

Corresponding author: Vito E. Cambria (vitoemanuele.cambria@phd.unipd.it)

Academic editor: Florian Jansen ♦ **Received** 15 November 2019 ♦ **Accepted** 31 January 2020 ♦ **Published** 4 May 2020

Abstract

Aim: To propose a Finite Mixture Model (FMM) as an additional approach for classifying large datasets of georeferenced vegetation plots from complex vegetation systems. **Study area**: The Italian peninsula including the two main islands (Sicily and Sardinia), but excluding the Alps and the Po plain. **Methods**: We used a database of 5,593 georeferenced plots and 1,586 vascular species of forest vegetation, created in TURBOVEG by storing published and unpublished phytosociological plots collected over the last 30 years. The plots were classified according to species composition and environmental variables using a FMM. Classification results were compared with those obtained by TWINSPAN algorithm. Groups were characterized in terms of ecological parameters, dominant and diagnostic species using the fidelity coefficient. Interpretation of resulting forest vegetation types was supported by a predictive map, produced using discriminant functions on environmental predictors, and by a non-metric multidimensional scaling ordination. **Results**: FMM clustering obtained 24 groups that were compared with those from TWINSPAN, and similarities were found only at a higher classification level corresponding to the main orders of the Italian broadleaf forest vegetation: *Fagetalia sylvaticae, Carpinetalia betuli, Quercetalia pubescenti-petraeae* and *Quercetalia ilicis*. At lower syntaxonomic level, these 24 groups were referred to alliances and sub-alliances. **Conclusions**: Despite a greater computational complexity, FMM appears to be an effective alternative to the traditional classification methods through the incorporation of modelling in the classificatory process. This allows classification of both the co-occurrence of species and environmental factors so that groups are identified not only on their species composition, as in the case of TWINSPAN, but also on their specific environmental niche.

Taxonomic reference: Conti et al. (2005).

Abbreviations: CLM = Community-level models; FMM = Finite Mixture Model; NMDS = non-metric multidimensional scaling.

Keywords

cluster analysis, finite mixture model, forest vegetation, Italian peninsula, vegetation plots

Introduction

The analysis of the spatial distribution of assemblages of communities is receiving increasing attention by ecologists (Nieto-Lugilde et al. 2017). To this purpose community-level models (CLM) are being used more and more, based on an "assemble-and-predict-together" strategy to simultaneously model multiple co-occurring species within a single process (Ferrier and Guisan 2006). They include methods that model the distribution of multiple species using a common set of environmental variables (De'ath 2002; Yee 2004, 2006; Leathwick et al. 2006). This feature makes CLM particularly promising for the classification of vegetation since the identification of one type is based on both its species composition and the environmental space it occupies (De Cáceres et al. 2015; Guarino et al. 2018).

Approaches to CLM clustering can be either based on minimizing a given loss function (for instance, the sum of within-group deviance), or can be based on associating each group to a specific joint density, which is parametrically specified. In this last case, CLM based clustering arises. While in standard (either hard or fuzzy) partitioning groups are summarized or represented by prototypes, in CLM clustering groups are represented by specific shapes of the corresponding probability density. Using such an approach, vegetation plots can be classified using the posterior probability that each belongs to a given component of the mixture, each component describing a group. Moreover, when the dataset is large, hierarchical approaches, based on the calculation of the pairwise (between plots) distances, rapidly become unfeasible. In this case, partitioning around prototypes, either means, medians or other, in a hard or a fuzzy perspective are usually adopted. However, much of these are based on simple Euclidean distances between each plot and the group prototypes that do not consider the dependence, the association and the covariance between the variables (plant species abundance values) characterizing the plots. In this respect, finite mixtures of multivariate Gaussian densities provide a simple, model-based, extension to the K-means method, allowing for overlapping clusters oriented according to the group-specific covariances and providing, *a posteriori*, for the classification of each plot to one of the groups. For this reason, among CLMs, Finite Mixture Modelling (FMM) is an emerging method and has already been used to identify marine bioregions on the Western Australian continental margin (Woolley et al. 2013) and forest physiognomic types in Italy (Attorre et al. 2014). In this latter paper, data from a National Forest Inventory were used, while here we test the applicability of FMM as a classification method for the forest vegetation of the Italian peninsula (including the major islands but excluding the Alps and the Po Plain). This area is characterized by great biogeographical and environmental variability and hosts a number of forest vegetation types, for which several classification schemes have been proposed (Pedrotti 1995; Pignatti 1998; Ubaldi 2003; Biondi et al. 2014; Mucina et al. 2016). The Italian peninsula is a broad ecotone between the Temperate and the Mediterranean regions (Attorre et al. 2014; Pesaresi et al. 2014). Boundaries between communities are not clearly defined having many species with overlapping ranges. Geo-pedological diversity, a variety of microclimates (Attorre et al. 2007), and a long history of disturbance that dates thousands of years and includes logging, fire, grazing, and plantation activities (Médail and Quézel 1999; Scarascia-Mugnozza et al. 2000; Vallejo et al. 2005), make the identification and classification of vegetation types difficult.

Within this framework, this paper aims to verify the applicability of FMM as classification method of vegetation plots using a complex case study and a large dataset, comparing the classification results with (1) those obtained by the TWINSPAN algorithm and (2) with current syntaxonomic classification schemes.

Methods

Data set

Observation data include 5,593 georeferenced vegetation plots of between 100 and 300 m^2 and 1,586 vascular species of forests in the Italian peninsula and major islands (Landucci et al. 2012; Agrillo et al. 2017). The database was created in TURBOVEG 3 (Hennekens and Schaminée 2001) by digitalizing and georeferencing published plots collected over the last 30 years (http://www.givd.info/ID/EU-IT-011).

Environmental covariates to be used in the statistical model were derived from a database with a spatial resolution of 1×1 km (Attorre et al. 2007): mean annual temperature (MeanT), mean minimum temperature of the coldest month (MinT), mean maximum temperature of the hottest month (MaxT), sum of mean monthly precipitation over summer (Ps) and winter months (Pw), and total annual precipitation (Ptot). We also used slope (SLO), derived from the GTOPO30 digital elevation model, geographical coordinates and a simplified geological map, derived from the geological map of Italy at 1:1.250.000 scale provided by the Italian Institute for Environmental Protection and Research, incorporating five main substrata: volcanic, arenaceous, carbonatic, clayey, sandy and conglomeratic.

Data analysis

We used a FMM to cluster vegetation plots, based on the assumption that data originate from one of K potential groups, also referred to as components. Each group is identified by a component, and each component is completely characterized by a distribution with known parametric form and component-specific parameters. When a (multivariate) Gaussian density is used to describe the component-specific distribution of observed plant species

cover, the component is identified by a specific center, defined by the mean vector (as the observed values are on abundance scale, we may hypothesize that similar plots will be characterized by similar values of abundance of the same species), and a specific shape, summarized by the covariance matrix, which allows for varying dependence between cover values corresponding to different plant species for plots in that component. The groups (components) are defined as homogeneous in the sense that they include plots that show similar vegetation as described by the plant species cover. Therefore, the observed plots can be allocated to one of the groups by using a criterion associated with the proximity between plots and group centers. This criterion is based on the posterior probability that a plot comes from that group (component of the finite mixture). The sum of the posterior probabilities over the components for a given plot is equal to 1, meaning that the plot has a varying degree of membership to all clusters in the population. We usually allocate a plot to a given cluster by finding that for which the posterior probability is maximum. At the end of the grouping step, each group will be characterized by a weight defined as the mean of posterior probabilities and refers to the (relative) frequency of plots allocated to that group. These terms can be interpreted as (prior) probabilities that a generic plot is randomly drawn from a "population of plots" belonging to that group (component of the finite mixture). We propose to model these (prior) probabilities as a function of so-called auxiliary variables (see e.g. McLachlan and Peel 2000). Thus, for each plot, the probability that the plot belongs to a group is a function (through a multinomial logistic model) of environmental parameters, as well as of geographical information, represented by class membership of neighboring plots.

After estimating the parameter vectors for the component-specific densities describing observed abundance, and the prior probability models, we derived the updated posterior probabilities as the (normalized) product of the prior information (based on covariates) and the density for that specific component.

These two steps can be jointly performed within the same estimation algorithm (e.g. using Latent Gold software, see Haughton et al. 2009) using multiple maximization sub-steps; we first estimate the group centers and shapes, and, successively and conditionally on the previous results, we estimate the effect of observed covariates on the probability to belong to a given group. A further alternative is based on the so-called two-three-step procedures, see Vermunt (2010). Rather than defining the prior probability of belonging to a cluster as a function of both environmental and geographical variables, according to the latter approach we first estimated the FMM and then built up a model where cluster membership is a function of geographical and environmental variables, through a multinomial logit model. This may be of help when the approach we propose does present convergence issues, and it defines a viable alternative and an approximate approach to model cluster membership as a function of plot-specific geographical and environmental features. For a formal description of the FMM see Attorre et al. (2014).

In this paper, we adapted the FMM to account for a large data matrix, formed by 5,593 vegetation plots and 1,586 species whose percentage cover is recorded. In this case the direct application of a FMM would be difficult, since it would require the computation and inversion of a 1,586 * 1,586 covariance matrix, with a very sparse structure. Looking at the distribution of the number of species observed in each plot, we see that the corresponding median value is equal to 81; if we look at the distribution of the number of plots each species is present in, the median value is equal to 7. The outcome of this is that of 10,402,980 values in the abundance data matrix, we have 10,241,820 (i.e. 98.45%) null values. Thus, rather than applying a FMM to the observed matrix of percentage covers, we fitted this model to a derived matrix, defined by projecting the original data matrix onto the space spanned by the first 20 principal components of the original data matrix using an approximate method (see Baglama and Reichel 2005) for singular value decomposition (SVD) of the observed, sparse, data matrix, using the R package *irlba* (Baglama and Reichel 2019). The number of principal components has been chosen by looking at stability and robustness of the obtained partition; we have considered 5 to 40 eigenvalues and chosen 20 as the best balance between model fit and simplicity/robustness. After employing the sparse SVD, we have extracted the matrix **A** corresponding to the eigenvectors of the covariance matrix of the observed sparse data **X**. We defined the derived matrix **Y=XA** and applied to **Y** the FMM with some backfitting to check whether a too high (low) number of dimensions was considered. The FMM model parameters have been estimated using the *mclust* R package (Fraley et al. 2017).

The optimal number of forest groups (components) was obtained according to penalized likelihood criteria (AIC – Akaike 1973; BIC – Schwarz 1978; CAIC – Hurvich and Tsai 1989; AIC3 – Bodzogan 1994). For all criteria, the lower the value of the index the better (more parsimonious) the fit to the observed data.

FMM classification was compared with that obtained by TWINSPAN (Hill 1979). The modified version of TWINSPAN (Roleček et al. 2009), implemented in JUICE (Tichý 2002), was used. This version, which has already been used in several comparative analyses of classification methods (Gauch and Whittaker 1981; Cao et al. 1997; Moss et al. 1999), allowed us to select the same number of groups obtained with the FMM classification. TWINSPAN pseudospecies cut levels for species abundance were set to 0-5-25 percentage scale units and five levels of divisions were chosen.

The obtained groups were characterized according to environmental parameters and diagnostic species, which were determined using the fidelity coefficient (phi) of Tichý and Chytrý (2006). To avoid phi being dependent on the size of the target site group, group size was standardized to equal the average size of all groups present in the data set (Tichý and Chytrý 2006). The phi values vary

independently of the concentration of species occurrence in the plots of individual groups. Statistical significance was obtained by a simultaneous calculation of Fisher's exact test. Species with phi values higher than 0.5 and Fisher's exact test significance lower than 0.001 were deemed to be diagnostic.

Interpretation of groups was supported by Kruskal's non-metric multidimensional scaling (NMDS) ordination (function isoMDS in the MASS R package, Venables and Ripley 2002). Moreover, we produced a predictive map by calculating discriminant functions based on environmental parameters that best discriminate between the estimated groups. These discriminant functions were used, post-estimation, to allocate (to groups) those plots from study areas where no information on plant species covers was available, while covariates describing environmental parameters were derived from available databases at a given resolution. The discriminant functions were estimated using the function *mda* from the R library. A confusion matrix of omission and commission errors was then calculated to evaluate the capacity of environmental factors to discriminate the groups obtained by FMM.

FMM R code and R libraries used for the statistical analyses are included in Suppl. material 1.

Results

FMM identified 24 groups, which were considered optimal according to all penalized likelihood criteria. However, four of these were discarded because they were characterized by few plots (less than 50), and they were quite heterogeneous. Descriptions of their environmental parameters, spatial distribution and syntaxonomic correspondences is presented in Suppl. material 2, while Suppl. material 3 shows mean and standard deviation of environmental parameters and dominant and diagnostic species of each group. With the support of the NMDS result (Figure 1) four main clusters were identified, corresponding to temperate beech forests (A), temperate chestnut-hornbeam forests (B) sub-Mediterranean deciduous forests (C) and evergreen Mediterranean forests (D). The distribution of classified vegetation plots is reported in Suppl. material 4, while the predictive distribution of groups and clusters is shown respectively in Figures 2 and 3.

Cluster A includes groups 8, 2, and 23. The first three can be found in temperate areas at an average altitude greater than 1000 m and are characterized by the dominance of *Fagus sylvatica* in groups 8 and 2, and by the codominance of this species with *Abies alba* in group 23 (Suppl. material 2 and 3). Group 8 is potentially widespread at the highest altitude along the Apennine chain and on the Etna volcano, while at a lower altitude, group 2 is mainly found in the southern part of the peninsula, and group 23 in the central-north (Figure 2). Cluster B includes only group 18, which is co-dominated by *Fagus sylvatica*, *Castanea sativa* and *Carpinus betulus,* with a

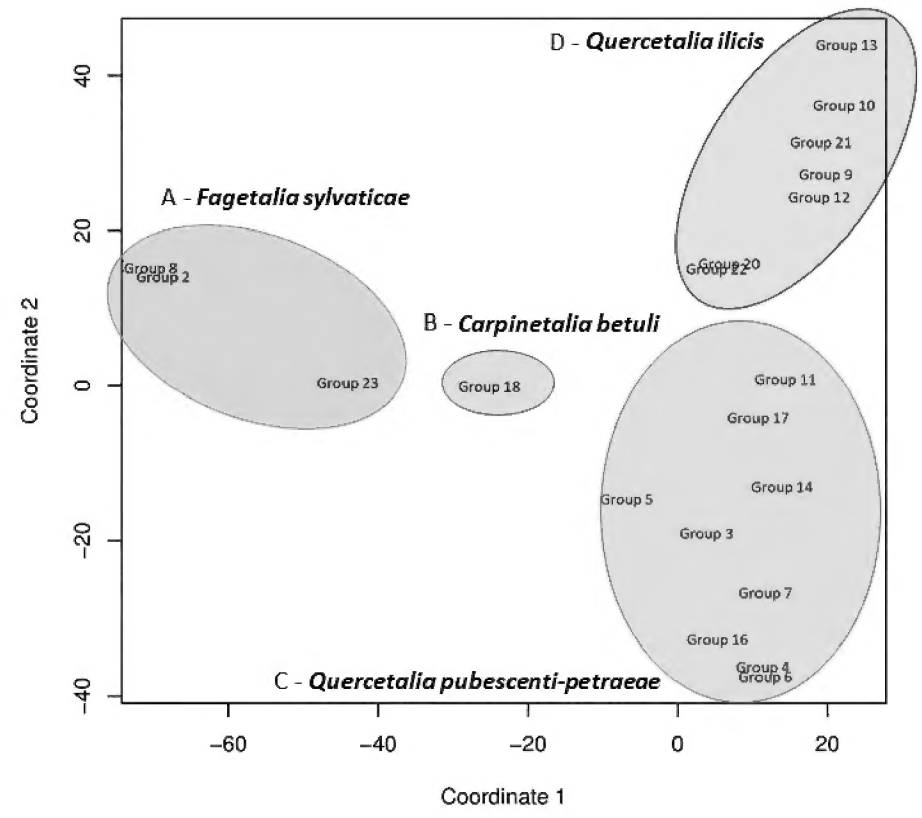

Figure 1. Kruskal's NMDS ordination of the vegetation groups. Due to the high number of plots only the centroids of the groups are shown. Stress values of the two components are 0.32 and 0.18, respectively.

distribution mainly localized in central Italy. Cluster C includes the sub-Mediterranean forests characterized by a high frequency of *Quercus cerris* in all groups, which can be accompanied by other deciduous tree species such as *Ostrya carpinifolia*, *Quercus pubescens* s.l., *Quercus frainetto* and *Fraxinus ornus* (Suppl. material 2 and 3). These groups occupy larger potential areas within an average altitudinal range from the coastal area up to 1000 m a.sl. Some of these can be very localized, such as group 4, characterized by forest stands dominated by *Quercus cerris* in the sub mountain areas of Liguria and northern Tuscany, and group 3, which is characterized by the codominance of *Quercus cerris* and *Ostrya carpinifolia* and is scattered throughout the peninsula (Figure 2). Others are quite widespread such as group 7 characterized by a mixed forest of *Quercus cerris* and *Quercus pubescens*, often with a dominated tree layer of *Carpinus orientalis* and *Erica arborea* and a potential distribution of about 28,000 km^2 mainly in central and southern Italy (Figure 2). Cluster D includes groups 20 and 22 characterized by the dominance of *Quercus suber*. Group 20 is localized in southern Italy and Sicily, while group 22 is potentially distributed in Sardinia and along the Tyrrhenian coast of the peninsula. Other groups within the cluster comprise formations dominated by *Quercus ilex* (Groups 9, 10, 12, 13 and 21). They can be subdivided into two main types: the first one mainly co-dominated by evergreen species at a lower altitude along the coast (Groups 13 and 21) and the second with deciduous tree species such as *Fraxinus ornus*, *Quercus frainetto*, *Quercus pubescens* and *Ostrya carpinifolia,* mainly localized in the inner part of the study area (Groups 9, 10 and 12).

TWINSPAN classification identified three main clusters, dominated by temperate broadleaved deciduous

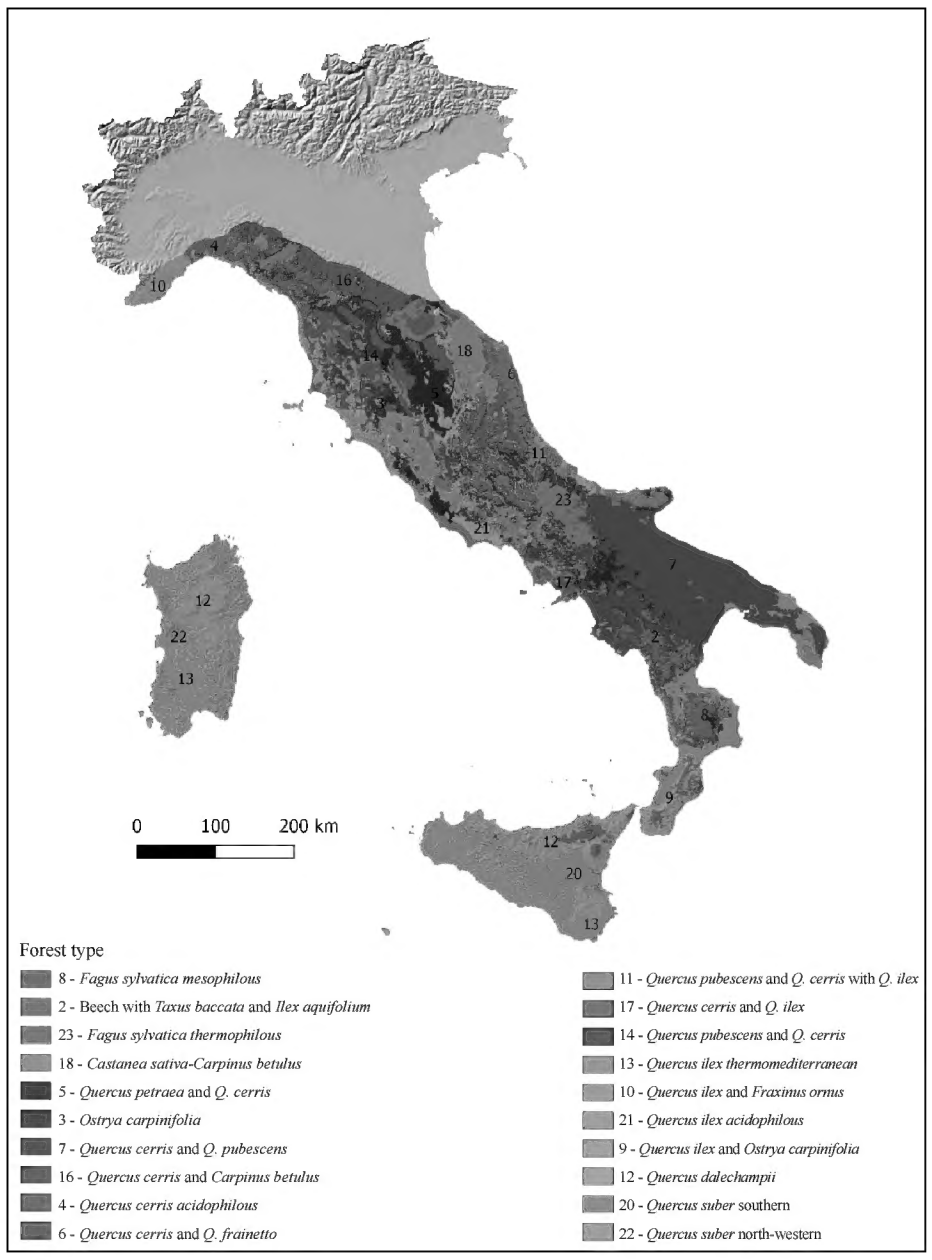

Figure 2. Map of the predictive distribution of the 20 groups based on the discriminant functions applied to environmental factors. The grey color indicates the part of the Italian peninsula not included in the analysis (Alps and the Po Plain).

Figure 3. Map of the potential distribution of the 4 clusters corresponding to main syntaxonomic forest orders recognized for Italy: A – *Fagetalia sylvaticae*, B – *Carpinetalia betuli*, C – *Quercetalia pubescenti-petraeae*, D – *Quercetalia ilicis*.

forests generally dominated by *Fagus sylvatica* (Groups 1–13), evergreen Mediterranean forests dominated by *Quercus suber* and *Quercus ilex* (Groups 14–18) and sub-Mediterranean deciduous forests dominated by *Quercus cerris* (Groups 19–24). The first TWINSPAN cluster corresponds to the four groups of the FMM classification (FMM groups 2, 8, 18 and 23, Table 1). The second cluster includes FMM Mediterranean evergreen groups, clearly differentiating *Quercus suber* and *Quercus ilex* dominated forests. The third TWINSPAN cluster contains all the sub-Mediterranean deciduous forest groups obtained with the FMM classification, group 24 accounting for more than 1000 of these plots.

The confusion matrix built to compare classified versus predicted plots highlighted that, with only some exceptions, environmental factors alone are insufficient to clearly discriminate among the groups identified by the FMM classification (Suppl. material 5). However, a significant difference emerges among clusters: beech forests (Group 2, 8, 23) appear to be better distinguishable, as indicated by the lower omission and commission errors. They are followed by evergreen Mediterranean forests. The poorest results were obtained for sub-Mediterranean deciduous forest types.

Discussion

The choice of an algorithm for the classification of vegetation plots depends on the objective of the classification and each algorithm has advantages and drawbacks (De Cáceres et al. 2015). The results of a classification algorithm can be evaluated by comparison with those of another and with current scientific knowledge on the vegetation type analyzed. In our study, the comparative analysis of FMM and TWINSPAN results highlighted good correspondence only at a high classification level where temperate, deciduous sub-Mediterranean and evergreen Mediterranean forest vegetation clusters were identified (Table 1). At lower levels, significant differences emerged with FMM classification producing groups with an even distribution of plots. Conversely, TWINSPAN split the homogeneous beech forests into many groups but identified two (groups 18 and 24) with 1000 plots each, including almost all the evergreen Mediterranean *Quercus ilex* dominated forests and the sub-Mediterranean deciduous forest dominated by *Quercus cerris* (Figure 4).

Consequently, FMM appears an effective alternative to traditional classification methods, such as TWINSPAN, to support the analysis of complex vegetation systems due to the ability to integrate both species composition and environmental factors into the modelled classificatory process.

Table 1. Comparative matrix between the 24 groups obtained by Finite Mixture Model classification (rows) and the 24 groups by the modified version of TWINSPAN (columns). Colors of the margins (groups) indicate membership to the clusters. Within the matrix, the red color indicates no correspondence among the groups. An increasing correspondence is highlighted by a color gradient from yellow to dark green.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Tot
2	1	5	50	15	12	32	9	21	79	44	4	10	80	0	0	0	0	0	0	0	0	0	0	1	363
8	4	10	19	142	34	18	106	21	47	95	0	0	0	0	0	0	0	0	0	1	0	0	0	0	497
23	0	12	47	28	4	12	3	0	24	35	28	51	132	0	0	0	0	0	2	0	0	1	0	5	384
18	0	3	11	0	6	4	0	0	5	0	30	21	79	0	0	0	2	0	10	0	0	3	1	12	187
3	1	0	0	0	0	0	0	0	0	0	79	42	13	0	0	0	1	0	3	0	1	8	0	50	198
4	0	0	0	0	0	0	0	0	0	0	17	0	1	1	0	0	0	3	11	0	1	10	3	50	97
5	0	0	5	0	0	1	0	0	1	0	27	9	25	0	0	0	0	10	35	0	5	1	1	61	181
6	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	18	0	7	6	1	160	197
7	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	1	15	7	13	1	8	4	9	98	163
11	0	0	0	0	0	0	0	0	0	0	7	3	0	0	0	0	18	59	3	0	0	9	0	205	304
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	39	0	0	0	14	3	323	397
16	1	0	0	0	0	0	0	0	1	0	84	23	81	0	0	0	0	0	5	0	3	2	0	38	238
17	0	0	0	0	0	1	0	0	0	0	6	1	2	7	0	3	11	43	43	0	13	7	3	55	195
9	0	0	0	0	0	0	0	0	0	0	3	2	3	0	0	0	21	31	5	0	22	14	1	11	113
10	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	87	165	4	1	0	26	0	46	333
12	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	7	237	38	2	0	1	16	66	13	381
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	55	390	0	1	0	0	0	2	449
21	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	28	36	246	0	0	0	1	1	16	357
20	0	0	0	0	0	0	0	0	0	0	0	0	0	18	15	76	19	33	0	1	0	0	2	2	166
22	0	0	0	0	0	0	0	0	0	0	0	0	0	96	44	46	2	56	0	0	0	0	0	0	244
1	1	9	0	0	2	0	0	0	0	0	8	0	0	4	0	0	0	4	16	0	7	0	0	5	56
15	0	0	0	0	0	0	0	0	0	0	0	10	9	0	0	0	0	0	2	0	0	1	0	9	31
19	0	0	0	0	0	0	0	0	0	0	1	3	3	0	0	0	4	0	6	0	1	3	2	10	33
24	0	0	1	0	0	0	0	0	0	0	1	0	1	2	0	2	0	0	7	0	2	1	0	12	29
Tot	8	39	133	185	58	68	118	42	157	174	308	179	429	157	59	164	526	1124	185	5	71	127	93	1184	5593

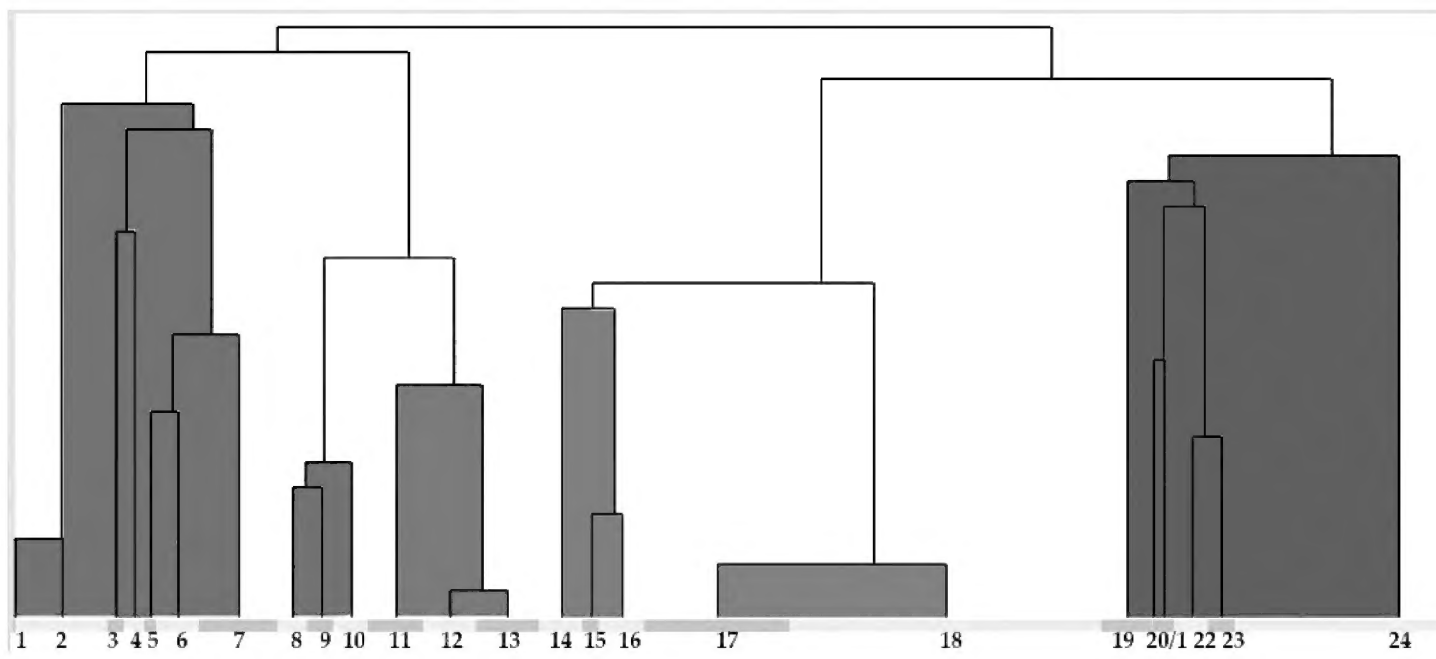

Figure 4. Modified TWINSPAN classification with 24 groups. Light blue color indicates groups belonging to *Fagetalia sylvaticae* (Groups 1–10), purple to *Carpinetalia betuli* (Groups 11–13), orange and red to *Quercetalia ilicis* (Groups 14–18), and green to *Quercetalia pubescenti-petraea* (Groups 19–24).

Moreover, since FMM identifies groups according to their ecological space, a predictive distribution map can also be produced (Figure 2) that better highlights geographic patterns than by viewing the distribution of classified plots alone (Suppl. material 4).

When compared with current syntaxonomic knowledge, the groups obtained by the FMM classification largely corresponded to several alliances and suballiances recognized for Italy according to Mucina et al. (2016) (Table 2). The environmental niche of groups also aligns well with that proposed in the relevant literature, while the floristic composition and the spatial distribution of groups can significantly differ. For instance, in Italy the temperate deciduous forest vegetation characterized by *Fagus sylvatica* and *Quercus* sp. pl. has been traditionally classified on the basis of a distinction between northern and southern syntaxa (see Blasi et al. 2004). This tradition began with Gentile (1970) in the study of beech forests of the Apennines, and was based on the recognition of a number of vicariant closely related species: *Geranium nodosum* (North) / *Geranium versicolor* (South), *Digitalis lutea* (N) / *Digitalis micrantha* (= *D. lutea* subsp. *australis*) (S), *Teucrium scorodonia* (N) / *Teucrium siculum* (S). This phytogeographical distinction was related to a sharp bioclimatic boundary between northern and southern Apennines, the former with no or limited summer drought stress and thus broadly referable to a temperate climate, and the latter with a more pronounced drought stress and

Table 2. Correspondence between the FMM group and the syntaxonomy in Mucina et al. (2016). The alliances are sorted according to an environmental gradient from temperate mesophilous to Mediterranean xeric.

FMM Group	Alliance in Mucina et al. (2016)
2	New alliance?
8	FAG-02B *Fagion sylvaticae* Luquet 1926
23	FAG-02C *Geranio striati-Fagion* Gentile 1970
18	FAG-03 *Carpinetalia betuli* P. Fukarek 1968
3	PUB-01F *Fraxino orni-Ostryion* Tomazic 1940
4	FAG-03C *Erythronio-Carpinion* (Horvat 1958) Marincek in Wallnofer et al. 1993
5	PUB-01L *Crataego laevigatae-Quercion cerridis* Arrigoni 1997
6	PUB-01L *Crataego laevigatae-Quercion cerridis* Arrigoni 1997
7	PUB-01L *Crataego laevigatae-Quercion cerridis* Arrigoni 1997
11	PUB-01G *Carpinion orientalis* Horvat 1958
14	PUB-01G *Carpinion orientalis* Horvat 1958
16	FAG-03C *Erythronio-Carpinion* (Horvat 1958) Marincek in Wallnofer et al. 1993
17	PUB-01L *Crataego laevigatae-Quercion cerridis* Arrigoni 1997
9	PUB-01M *Pino calabricae-Quercion congestae* S. Brullo et al. 1999
10	QUI-01D *Fraxino orni-Quercion ilicis* Biondi, Casavecchia et Gigante in Biondi et al. 2013
12	PUB-01M *Pino calabricae-Quercion congestae* S. Brullo et al. 1999
13	QUI-01A *Quercion ilicis* Br.-Bl. ex Molinier 1934
20	QUI-01E *Erico-Quercion ilicis* S. Brullo et al. 1977
21	QUI-01E *Erico-Quercion ilicis* S. Brullo et al. 1977
22	QUI-01E *Erico-Quercion ilicis* S. Brullo et al. 1977

thus referable to a sub-Mediterranean or supra-Mediterranean climate (Feoli and Lagonegro 1982; Pignatti and Wikus Pignatti 1990). This led to the definition of northern and southern alliances or suballiances, for instance, the northern *Geranio nodosi-Fagion* (= *Cardamino kitaibelii-Fagenion*) and southern *Geranio versicoloris-Fagion* (= *Geranio striati-Fagenion*) (Feoli and Lagonegro 1982; for the nomenclature see Di Pietro et al. 2004).

In our analysis, a more complex pattern emerged: the gradient of different bioclimates, from temperate to sub-Mediterranean, with decreasing water availability and increasing temperature, follows not only the phytogeographical sector but also an altitudinal gradient. For instance, temperate beech forests of the upper altitude are potentially distributed all along the peninsula including the Etna volcano in Sicily (Group 8), while lower altitude beech forests (Groups 2 and 23) are distributed respectively in the south and in the central north (Figure 2). This result substantially agrees with Willner et al. (2017), even though the geographic boundaries between groups 2 and 23 are different because the southern group is more localized than indicated by Willner et al. (2017). Moreover, high altitude beech forests (Group 8) are floristically relatively different from the currently recognized alliances since they include many local endemics from both the north and south Apennines (Suppl. material 3).

Cluster B includes only group 18 and can be referred to the *Carpinetalia betuli* order (Mucina et al. 2016), which is united with *Fagetalia sylvaticae* in the *Carpino-Fagetea* class. NMDS analysis (Fig 1) confirmed its floristic affinity with the beech forests, even though in the Italian peninsula it is spatially and ecologically embedded within the deciduous sub-Mediterranean forests (Suppl. material 3 and Figure 3).

Sub-Mediterranean deciduous oak forests of cluster C are characterized by a complex geographic pattern along the Apennines, which cannot be explained only by the combination of geo-climatic factors, as is highlighted by the very high omission errors of the confusion matrix (Suppl. material 5). These groups show a good correspondence with many alliances and sub-alliances reported in the prodrome of the Italian vegetation (Biondi et al. 2014). Nonetheless, in our study, some syntaxonomic units are split into two or more floristically and ecologically well-defined groups. For instance, the *Cratego laevigati-Quercion cerridis* alliance is split into two groups (6 and 7), with different floristic composition and distinct ecology. Another similarity is represented by the alliance *Carpinion orientalis*, for which three suballiances, *Laburno anagyroidis-Ostryenion, Cytiso sessilifolii-Quercenion pubescentis* and *Lauro nobilis-Quercenion pubescentis* have been identified for Italy (Blasi et al. 2004). In our analysis, they correspond respectively to groups 3, 11 and 14. However, a comprehensive comparison with the *Carpinion orientalis* of the Balkans is still lacking, as well as with the *Quercion pubescenti-petraeae* described by Braun-Blanquet for Provence and Catalonia, which seems very similar to group 14 and to which this has been sometimes referred to (Ubaldi 2003).

The geographic pattern also characterizes the evergreen Mediterranean forests, which are difficult to classify due to the low number of characteristic species, especially in the herbaceous layer. FMM (and also TWINSPAN, see Table 1) clearly differentiated *Quercus suber* and *Quercus ilex* dominated forest vegetation (Figure 1). The former includes groups 20 and 22, one distributed in southern Italy, and the other one in Sardinia and the northern Tyrrhenian coast. A geographic pattern is also evident for evergreen forests dominated by *Quercus ilex*: group 10 is mainly distributed in Liguria and central Italy, group 13 mainly in southern Sardinia and Sicily in the thermo-Mediterranean region, and group 21 includes the coastal forests along both sides of the Italian peninsula. Mixed evergreen and deciduous forests are localized in the supra-Mediterranean region respectively, group 12, co-dominated by *Quercus pubescens* s.l., in Sicily and

Sardinia, and group 9 co-dominated by *Quercus frainetto* very localized in central Italy and Calabria. This classification significantly differs from that currently indicated in the Italian vegetation prodrome, which for the evergreen Mediterranean forests in Italy recognizes only four sub-alliances (Biondi et al. 2003; Bacchetta et al. 2004; Brullo et al. 2008). In our study, we find instead seven groups, which are not exceedingly well characterized from a floristic point of view (even though it must be taken into account the floristic poverty of the *Quercetea ilicis* forests) but are instead perfectly reasonable under an ecological and phytogeographical point of view. For instance, an interesting distinction of both *Quercus suber* and *Quercus ilex* forests in northern-central (22 for *Quercus suber*, and 10 for *Quercus ilex*) and southern groups (20 for *Quercus suber* and 13 for *Quercus ilex*) can be observed. This result has important phytogeographical and syntaxonomic implications that are related to the limits between the meso-Mediterranean and thermo-Mediterranean regions, and it deserves a broader analysis at the continental scale.

The 20 groups can be aggregated in four clusters corresponding to the main syntaxonomic orders recognized for the Italian peninsula: *Carpinetalia betuli, Fagetalia sylvaticae*, *Quercetalia ilicis Quercetalia pubescenti-petraeae* (Figure 3). Their spatial distribution also largely corresponds to the bioclimates recognized by Rivas Martínez for Italy (Rivas Martínez et al. 2004), even though the boundary of the sub-Mediterranean region shifted more south especially in the Apulia region. The bioclimatic limit defined by Rivas Martínez has a better correspondence with the results by Bohn et al. (2003) and Attorre et al. (2014). However, these authors based their biogeographical analyses only on dominant tree species, while in our analysis we included the whole species composition of forest vegetation plots. This also explains why Sardinia is completely classified as *Quercetalia ilicis*, whereas in the previous studies patches of sub-Mediterranean forest vegetation, characterized by stands co-dominated by *Quercus pubescens* s.l. and *Quercus ilex,* were recognized.

Conclusion

Despite a greater computational complexity, Finite Mixture Model seems to be a promising classificatory approach when dealing with the analysis of complex vegetation systems and using a large dataset. This relied on the possibility of modelling in the classification process both the co-occurrence of species and environmental variables so that groups are identified not only based on their species composition, such as in the case of TWINSPAN, but also on their specific environmental niche. These features can effectively highlight geographical patterns as depicted by predictive maps and support the interpretation of classification results.

Data availability

Primary data are stored in the European Vegetation Archive (Chytrý et al. 2016) and the Sapienza University vegetation database (https://www.givd.info/ID/EU-IT-011).

Author contributions

F.A., V.E.C., E.A. and G.F. conceived the study, M.A. and L.M. run the statistical analyses, and N.A., M.D.S., T.S., R.G., C.M., M.M. and F.S. contributed to the interpretation of results.

Acknowledgements

We would like to thank Laura Clarke for revising the text and all those who collected vegetation-plot data in the field and integrated these data in the Sapienza database (https://www.givd.info/ID/EU-IT-011).

References

Agrillo E, Alessi N, Massimi M, Spada F, De Sanctis M, Francesconi F, Cambria VE, Attorre F (2017) Nationwide Vegetation Plot Database - Sapienza University of Rome: State of the art, basic figures and future perspectives. Phytocoenologia 47: 221–229. https://doi.org/10.1127/phyto/2017/0139

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (Eds) International Symposium on Information Theory. Akademiai Kiado, Budapest, HU, 1033–1055.

Attorre F, Alfò M, De Sanctis M, Francesconi F, Bruno F (2007) Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. International Journal of Climatology 27: 1825–1843. https://doi.org/10.1002/joc.1495

Attorre F, Francesconi F, De Sanctis M, Alfò M, Martella F, Valenti R, Vitale M (2014) Classifying and Mapping Potential Distribution of Forest Types Using a Finite Mixture Model. Folia Geobotanica 49: 313–335. https://doi.org/10.1007/s12224-012-9139-8

Bacchetta G, Bagella S, Biondi E, Farris E, Filigheddu R, Mossa L (2004) A contribution to the knowledge of the order Quercetalia ilicis Br.-Bl. ex Molinier 1934 of Sardinia. Fitosociologia 41: 29–51.

Baglama J, Reichel L (2005) Augmented implicitly restarted lanczos bidiagonalization methods. SIAM Journal on Scientific Computing 27: 19–42. https://doi.org/10.1137/04060593X

Baglama J, Reichel L (2019) Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices, R package version 2.3.3. https://CRAN.R-project.org/package=irlba.

Biondi E, Casavecchia S, Gigante D (2003) Contribution to the syntaxonomic knowledge of the Quercus ilex L. woods of the Central European Mediterranean Basin. Fitosociologia 40: 129–156.

Biondi E, Blasi C, Allegrezza M, Anzellotti I, Azzella MM, Carli E, Casavecchia S, Copiz R, Del Vico E, … Zivkovic L (2014) Plant communities of Italy: The Vegetation Prodrome. Plant Biosystems 148: 728–814. https://doi.org/10.1080/11263504.2014.948527

Blasi C, Di Pietro R, Filesi L (2004) Syntaxonomical revision of Quercetalia pubescenti-petraeae in the Italian peninsula. Fitosociologia 41: 87–164.

Bohn U, Neuhäuslm R, Gollub G, Hettwer C, Neuhäuslová Z, Raus T, Schlüter H, Weber H (2003) Map of the natural vegetation of Europe. German Federal Agency for Nature Conservation, Bonn, DE.

Bodzogan H (1994) Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In: Bodzogan H (Ed.) Proceedings of the first US/Japan conference on the frontiers of statistical modelling: An informational approach. Multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, NL, 69–113. https://doi.org/10.1007/978-94-011-0800-3_3

Brullo S, Gianguzzi L, La Mantia A, Siracusa G (2008) La classe *Quercetea ilicis* in Sicilia. Bollettino dell'Accademia Gioenia di Scienze Naturali 41: 1–124.

Cao Y, Bark AW, Williams WP (1997) A comparison of clustering methods for river benthic community analysis. Hydrobiologia 347: 25–40. https://doi.org/10.1023/A:1002938721135

Chytrý M, Hennekens SM, Jiménez-Alfaro B, Knollová I, Dengler J, Jansen F, Landucci F, Schaminée JH, Aćić S, … Yamalov S (2016) European Vegetation Archive (EVA): an integrated database of European vegetation plots. Applied Vegetation Science 19: 173–180. https://doi.org/10.1111/avsc.12191

Conti F, Abbate G, Alessandrini A, Blasi C (2005) An annotated checklist of the Italian vascular flora. Palombi Editori, Roma, IT.

De'ath G (2002) Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83: 1105–1117. https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2

De Cáceres M, Chytrý M, Agrillo E, Attorre F, Botta-Dukát Z, Capelo J, Czúcz B, Dengler J, Ewald E, … Wiser SK (2015) A comparative framework for broad-scale plot-based vegetation classification. Applied Vegetation Science 18: 543–560. https://doi.org/10.1111/avsc.12179

De Cáceres M, Franklin SB, Hunter JT, Landucci F, Dengler J, Roberts DW (2018) Global overview of plot-based vegetation classification approaches. Phytocoenologia 48: 101–112. https://doi.org/10.1127/phyto/2018/0256

Di Pietro R, Izco J, Blasi C (2004) Contribution to the nomenclatural knowledge of *Fagus sylvatica* woodlands of southern Italy. Plant Biosystems 138: 27–36. https://doi.org/10.1080/11263500410001684099

Feoli E, Lagonegro M (1982) Syntaxonomical analysis of beech woods in the Apennines (Italy) using the program package IAHOPA. Vegetatio 50: 129–173. https://doi.org/10.1007/BF00364109

Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43: 393–404. https://doi.org/10.1111/j.1365-2664.2006.01149.x

Fraley C, Raftery AE, Scrucca L (2017) mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. R package version 5.4. https://CRAN.R-project.org/package=mclust

Gauch HG, Whittaker RH (1981) Hierarchical classification of community data. Journal of Ecology 69: 537–557. https://doi.org/10.2307/2259682

Gentile S (1970) Sui faggeti dell'Italia meridionale (Beech woodlands of Southern Apennines). Atti Dell'istituto Botanico dell'Università Di Pavia 65: 207–306.

Guarino R, Willner W, Pignatti S, Attorre F, Loidi JJ (2018) Spatio-temporal variations in the application of the Braun-Blanquet approach in Europe. Phytocoenologia 48: 239–250. https://doi.org/10.1127/phyto/2017/0181

Haughton D, Legrand P, Woolford S (2009) Review of Three Latent Class Cluster Analysis Packages: Latent GOLD, poLCA, and MCLUST. The American Statistician 63: 81–91. https://doi.org/10.1198/tast.2009.0016

Hennekens SM, Schaminée JHJ (2001) TURBOVEG, a comprehensive data base management system for vegetation data. Journal of Vegetation Science 12: 589–591. https://doi.org/10.2307/3237010

Hill MO (1979) TWINSPAN – A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca, NY, US.

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76: 297–307. https://doi.org/10.1093/biomet/76.2.297

Landucci F, Acosta ATR, Agrillo E, Attorre F, Biondi E, Cambria VE, Chiarucci A, Del Vico E, De Sanctis M, … Venanzoni R (2012) VegItaly: The Italian collaborative project for a national vegetation database. Plant Biosystems 146: 756–763. https://doi.org/10.1080/11263504.2012.740093

Leathwick JR, Elith J, Hastie T (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188–196. https://doi.org/10.1016/j.ecolmodel.2006.05.022

McLachlan G, Peel D (2000) Finite mixture models. In: Wiley J (Ed.) Probability and Statistics. Wiley Series, New York, US. https://doi.org/10.1002/0471721182

Médail F, Quézel P (1999) Biodiversity hotspots in the Mediterranean Basin: setting global conservation priorities. Conservation Biology 13: 1510–1513. https://doi.org/10.1046/j.1523-1739.1999.98467.x

Moss D, Wright JF, Furse MT Clarke RT (1999) A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. Freshwater Biology 41: 167–181. https://doi.org/10.1046/j.1365-2427.1999.00376.x

Mucina L, Bültmann H, Dierßen K, Theurillat JP, Raus T, Čarni A, Šumberová K, Willner W, Dengler J, … Tichý L (2016) Vegetation of Europe: hierarchical floristic classification system of vascular plant, bryophyte, lichen, and algal communities. Applied Vegetation Science 19: 3–264. https://doi.org/10.1111/avsc.12257

Nieto-Lugilde D, Maguire KC, Blois JL, Williams JW, Fitzpatrick MC (2017) Multiresponse algorithms for community-level modeling: review of theory, applications, and comparison to species distribution models. Methods in Ecology and Evolution 9: 834–848. https://doi.org/10.1111/2041-210X.12936

Pedrotti F (1995) La vegetazione forestale italiana. Atti Convegni Lincei 115: 39–78.

Pesaresi S, Galdenzi D, Biondi E, Casavecchia S (2014) Bioclimate of Italy: application of the worldwide bioclimatic classification system. Journal of Maps 10: 538–553. https://doi.org/10.1080/17445647.2014.891472

Pignatti S, Wikus Pignatti E (1990) Le cenosi a cerro e frainetto della penisola e della Sicilia. Notiziario Fitosociologico 23: 107–124.

Pignatti S (1998) I boschi d'Italia. UTET, Torino, IT.

Rivas Martínez S, Penas A, Díaz TE (2004) Bioclimatic and biogeographic maps of Europe – Bioclimates. Cartographic Service, University of Leon, ES. http://www.globalbioclimatics.org/form/bi_med.htm

Rodwell JS, Evans D, Schaminée JHJ (2018) Phytosociological relationships in European Union policy-related habitat classifications. Rendiconti Lincei 29: 237–249. https://doi.org/10.1007/s12210-018-0690-y

Roleček J, Tichý L, Zelený D, Chytrý M (2009) Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. Journal of Vegetation Science 20: 596–602. https://doi.org/10.1111/j.1654-1103.2009.01062.x

Scarascia-Mugnozza G, Oswald H, Piussi P, Radoglou K (2000) Forests of the Mediterranean region: gaps in knowledge and research needs. Forest Ecology and Management 132: 97–109. https://doi.org/10.1016/S0378-1127(00)00383-2

Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics 6: 461–464. https://doi.org/10.1214/aos/1176344136

Tichý L (2002) JUICE, software for vegetation classification. Journal of Vegetation Science 13: 451–453. https://doi.org/10.1111/j.1654-1103.2002.tb02069.x

Tichý L, Chytrý M (2006) Statistical determination of diagnostic species for site groups of unequal size. Journal of Vegetation Science 17: 809–818. https://doi.org/10.1111/j.1654-1103.2006.tb02504.x

Ubaldi D (2003) La vegetazione boschiva d'Italia: Manuale di Fitosociologia forestale. Clueb, Bologna, IT, 1–368.

Vallejo R, Aronson J, Pausas JG, Cortina J (2005) Restoration of Mediterranean woodlands. In: van Andel J, Aronson J (Eds) Restoration ecology: a European perspective. Blackwell Science, Oxford, GB, 193–207. https://doi.org/10.1002/9781118223130.ch11

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, Fourth edition. Springer, New York, US. https://doi.org/10.1007/978-0-387-21706-2

Vermunt JK (2010) Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis 18: 450–469. https://doi.org/10.1093/pan/mpq025

Willner W, Jiménez-Alfaro B, Agrillo E, Biurrun I, Campos JA, Čarni A, Casella L, Csiky J, Ćušterevska R, … Chytrý M (2017) Classification of European beech forests: a Gordian Knot? Applied Vegetation Science 20: 494–512. https://doi.org/10.1111/avsc.12299

Woolley SNC, McCallum AW, Wilson R, O'Hara TD, Dunstan PK (2013) Fathom out: biogeographical subdivision across the Western Australian continental margin – a multispecies modelling approach. Diversity and Distributions 19: 1506–1517. https://doi.org/10.1111/ddi.12119

Yee TW (2004) A new technique for maximum-likelihood canonical gaussian ordination. Ecological Monographs 74: 685–701. https://doi.org/10.1890/03-0078

Yee TW (2006) Constrained additive ordination. Ecology 87: 203–213. https://doi.org/10.1890/05-0283

E-mail and ORCID

Fabio Attorre (fabio.attorre@uniroma1.it), ORCID: http://orcid.org/0000-0002-7744-2195
Vito E. Cambria (Corresponding author, vitoemanuele.cambria@phd.unipd.it), ORCID: http://orcid.org/0000-0003-0481-6368
Emiliano Agrillo (emiliano.agrillo@isprambiente.it), ORCID: http://orcid.org/0000-0003-2346-8346
Nicola Alessi (nicola.alessi@natec.unibz.it)
Marco Alfò (marco.alfo@uniroma1.it), ORCID: http://orcid.org/0000-0001-7651-6052
Michele De Sanctis (michele.desanctis@uniroma1.it), ORCID: http://orcid.org/0000-0002-7280-6199
Luca Malatesta (luca.malatesta@uniroma1.it), ORCID: http://orcid.org/0000-0003-1887-4163
Tommaso Sitzia (tommaso.sitzia@unipd.it), ORCID: http://orcid.org/0000-0001-6221-4256
Riccardo Guarino (riccardo.guarino@unipa.it), ORCID: http://orcid.org/0000-0003-0106-9416
Corrado Marcenò (marceno.corrado@ehu.eus), ORCID: http://orcid.org/0000-0003-4361-5200
Marco Massimi (marco.massimi@hotmail.com)
Francesco Spada (francesco.spada@uniroma1.it)
Giuliano Fanelli (giuliano.fanelli@gmail.com), ORCID: http://orcid.org/0000-0002-3143-1212

Supplementary material

Supplementary material 1
MM R code and R libraries used for the statistical analyses (.R)
Link: https://doi.org/10.3897/VCS/2020/48518.suppl1

Supplementary material 2
Ecological, physiognomic and distributional features, floristic composition and syntaxonomy of groups (.DOCX)
Link: https://doi.org/10.3897/VCS/2020/48518.suppl2

Supplementary material 3
Ecological parameters, dominant and diagnostic species of the groups (.XLSX)
Link: https://doi.org/10.3897/VCS/2020/48518.suppl3

Supplementary material 4
Maps of the distribution of the classified plots of each group (.JPG)
Link: https://doi.org/10.3897/VCS/2020/48518.suppl4

Supplementary material 5
Confusion matrix generated for the accuracy assessment of the potential distribution map of groups (.DOCX)
Link: https://doi.org/10.3897/VCS/2020/48518.suppl5